Interspeech 2024
1-5 September 2024, Kos, Greece

Qifusion-Net: Layer-adapted Stream/Non-stream Model for End-to-End Multi-Accent Speech Recognition

Jinming Chen[1], *Jingyi Fang*[1], *Yuanzhong Zheng*[1], *Yaoxuan Wang*[1], *Haojun Fei*[1]

[1]Qifu Technology, China

chenjinming-jk@360shuke.com, fangjingyi-jk@360shuke.com, zhengyuanzhong-jk@360shuke.com, wangyaoxuan-jk@360shuke.com, zhangchulan-jk@360shuke.com

Abstract

Currently, end-to-end (E2E) speech recognition methods have achieved promising performance. However, auto speech recognition (ASR) models still face challenges in recognizing multi-accent speech accurately. We propose a layer-adapted fusion (LAF) model, called Qifusion-Net, which does not require any prior knowledge about the target accent. Based on dynamic chunk strategy, our approach enables streaming decoding and can extract frame-level acoustic feature, facilitating fine-grained information fusion. Experiment results demonstrate that our proposed methods outperform the baseline with relative reductions of 22.1% and 17.2% in character error rate (CER) across multi accent test datasets on KeSpeech and MagicData-RMAC.
Index Terms: multi-accent speech recognition, layer-adapted fusion, stream/non-stream decoding, cross-attention

1. Introduction

In recent years, end-to-end (E2E) speech recognition (ASR) has significantly benefited from the high-resource languages and increasing large model size [1, 2]. This improvement has enabled the effective mitigation of recognition degraded caused by various acoustic environments [3]. As a result, E2E ASR systems have found widely-used in commercial speech recognition products [4–7]. However, it is well known that the performance of even large ASR models degrades significantly when the speakers have varying degrees of accent pronunciation [8]. Accent is a special way of pronunciation, which is mainly influenced by regional culture, speaking style and the education level of the speaker [9]. For example, in the remote areas or villages of southern China, those accents are quite different from the pronunciation of Mandarin and seriously affects the recognition accuracy of the E2E ASR model [8]. To some extent, training a specific accent ASR model can solve the problem of recognition accuracy degradation. Achieving high recognition accuracy in multi-accent system without pre-accent category information has significant commercial value for the application of E2E ASR model.

Recently, adversarial learning [10, 11], transfer learning [12, 13], multi-tasking learning (MTL) [14, 15] and other deep learning methods have been developed greatly to eliminate the recognition bias due to the accent in the ASR task. The primary concept behind adversarial learning and transfer learning methods involves initially training the model on an extensive corpus of speech data, followed by fine-tuning it specifically on the accent dataset [12, 16, 17]. Good performance is often obtained in a single accent system. For multi-accent systems, it is often necessary to introduce additional information to guide for the different accents. The direct way to introduce accent information is to concatenate a one-hot accent vector into the input acoustic features [18]. Others, accent identification (AID) models are used to generate embeddings [19, 20]. For instance, the authors in [21, 22] suggest connecting accent embeddings and acoustic features to adapt the acoustic model. In [23], they utilized well-trained accent classifiers to extract accent embedding for layer-to-layer adaptation of E2E ASR models. A multi-task framework was proposed in [21, 24] to jointly model ASR and AID tasks. All previous researches have significantly enhanced the accuracy of accent speech recognition in specific contexts. However, thus far, there has been a lack of consideration for accent recognition in real-time streaming scenarios and the fusion of fine-grained accent information at the frame level, which holds greater practical applicability and reference significance for MTL.

In this paper, we aim at improving the recognition accuracy in a multi-accent system. Three contributions are explored: 1) A new fusion strategy Layer-adapted fuison (LAF) module is proposed to extract accent information in shared acoustic encoder. 2) Fine-grained fusion of accent information at frame-level is achieved through the utilization of a cross-attention module, which effectively eliminates the impact of accent on the acoustic model. 3) Based on the dynamic chunk strategy, the model realizes the unification of streaming and non-streaming decoding modes. The results demonstrate that both stream/non-stream Qifusion-Net achieve the highest accuracy for accented speech datasets. The CER demonstrates the relative decrease of 22.1% and 17.2% in the multi accent test datasets, compared to the baseline in KeSpeech and MagicData-RAMC.

The remaining sections of this paper are structured as follows: Section 2 introduces our Layer-adapted fusion module integrated with the MTL E2E ASR system. In Section 3, experimental results are presented and analyzed. Finally, Section 4 provides the conclusion.

2. Layer-adapted for E2E Multi-Accent ASR

The whole model architecture of the proposed layer-adapted for E2E multi-accent ASR is illustrated in Figure 1. In this section, we first give a brief review of a general acoustic encoder based on the convolution-augmented transformer (Conformer) model in Section 2.1. Then the proposed layer-adapted fusion (LAF) module is introduced in Section 2.2. The LAF module is designed for filtering the fine-grained accent information from the shared acoustic encoder with two different structures. Furthermore, we present the cross-attention module in Section 2.3, which integrates the acoustic and accent information. Finally, the MTL training of our methods and the stream/non-stream decoding modes are described in Section 2.4.

10.21437/Interspeech.2024-930

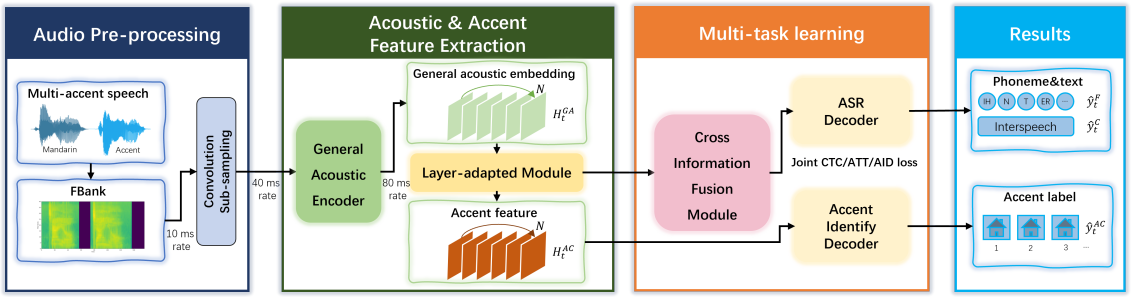
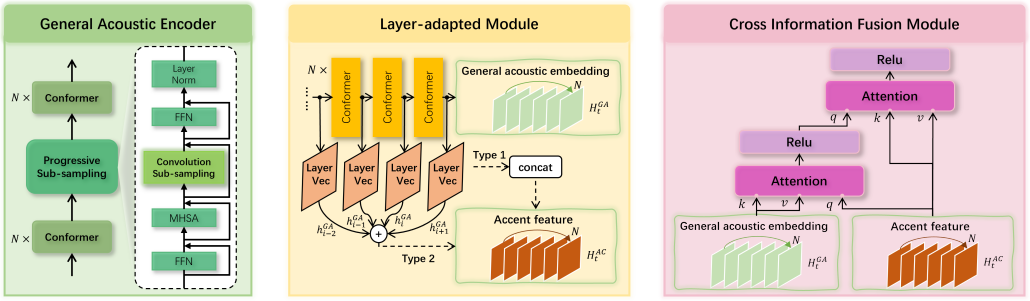


Figure 1: *Schematic architecture of the proposed layer-adapted for end-to-end multi-accent ASR model.*



(a) General Acoustic Encoder (b) Layer-adapted Module (c) Cross Information Fusion Module

Figure 2: *Key parts of the model architecture.*

2.1. Conformer-based Acoustic Encoder

The acoustic encoder structure innovations in this paper are based on the general conformer E2E ASR model. The Conformer model integrates convolution layers into the transformer module to enhance the local modeling capability of signal sequences. Given its exceptional performance in the E2E ASR task, we employ the conformer-based encoder depicted in Figure 2(a) to comprehensively extract frame-level acoustic features from incoming multi-accent speech.

The original conformer model is mainly composed of four modules: two feed-forward modules (FFN), a multi-head self-attention module (MHSA) and a convolution module in the middle. When the step length of speech frame is set to 10ms, there will be redundancy in the general acoustic features which will affect the fine-grained accent feature extraction to some extent. Hence a progressive sub-sampling operation is applied in the general acoustic encoder. The feature compression in time dimension will be conducive to the following effective accent features extraction by layer-adapted module in Section 2.2.

Given the FBank features X_t of the input multi-accent speech, the frame-level general acoustic features H_t^{GA} of the conformer block can be mathematically defined as follows:

$$h_t^{\mathrm{FFN}_1} = X_t + \frac{\mathrm{FFN}(X_t)}{2} \tag{1}$$

$$h_t^{\mathrm{MHSA}} = h_t^{\mathrm{FFN}_1} + \mathrm{MHSA}(h_t^{\mathrm{FFN}_1}) \tag{2}$$

$$h_t^{\mathrm{Conv}} = h_t^{\mathrm{MHSA}} + \mathrm{Conv}(h_t^{\mathrm{MHSA}}) \tag{3}$$

$$h_t^{\mathrm{FFN}_2} = h_t^{\mathrm{Conv}} + \frac{\mathrm{FFN}(h_t^{\mathrm{Conv}})}{2} \tag{4}$$

$$H_t^{GA} = \mathrm{LayerNorm}(h_t^{\mathrm{FFN}_2}) \tag{5}$$

On the one hand, this design can gradually reduce the dimension of input sequence, so that the global acoustic features can be projected to a wider dimension [25], and on the other hand, it can reduce the computational complexity of streaming mode.

2.2. Layer-adapted Module

Current solutions for multi-accent ASR task commonly adopt the acoustic features of a universal ASR model as the input to train an accent identifier (AID) model. However, such methods lack information sharing between the two models which leads to both performance degradation in a multi-task learning of ASR and AID. As shown in Figure 2(b), we proposed a layer-adapted module to extract fine-grained accent information from different layers of acoustic encoder while facilitating frame-by-frame correction of ASR results by cross-attention module in Section 2.3.

2.2.1. Adapted Layer

Numerous studies have indicated that distinct layers of the ASR encoder possess the capability to extract speech information at varying levels. As the depth of the ASR layers increases, a greater abundance of localized information becomes available. In this paper, we use the layer after the progressive sub-sampling operation of acoustic structure as adapted layers. In the training process, we introduce learnable adaptive weights that are multiplied with the adapted layers. Both concatenate and sum operations can be selected as layer-adapted connectivity options (Figure 2(b)).

Table 1: *Ablation and contrast experiments results of the AID task on the KeSpeech test dataset.*

ID	Model	AID ACC(%)
C1	KeSpeech Baseline	61.13
C2	Kaldi-xvector	56.34
C3	ResNet-34	61.13
C4	ECAPA-TDNN	60.77
C6	DIMNet w/ LM	78.57
L1	Qifusion-Net-L6	33.64
L2	Qifusion-Net-L7	34.67
L3	Qifusion-Net-L8	**39.75**
L4	Qifusion-Net-L9	33.08
L5	Qifusion-Net-L10	25.95
L6	Qifusion-Net-L11	26.17
Q2	Qifusion-Net-ns	**79.10**

2.2.2. Accent Identify Decoder

After extracting the fused accent features from the adapted layers in the acoustic encoder, we propose a two-layer causal convolutional structure and a linear-based discriminator to construct the AID. This module can effectively distill accent information and provides frame-by-frame classification of input multi-accent speech into different accent categories.

2.3. Cross-attention Module

It is widely acknowledged that the MTL approach can enhance the performance of each task by facilitating the sharing of feature information. As shown in Figure 2(c), we use the output H_t^{GA} of the general acoustic encoder as the key, and the accent embedding H_t^{AC} obtained in the layer adapted module as the query to carry out cross-information fusion. The frame-level accent embedding features contributes to eliminate the distortion of acoustic features caused by different degrees of accent pronunciation in a multi-accent speech in order to improve the accuracy of accent ASR.

The following shows calculation process of cross-information fusion based on attention mechanism:

$$Q_t = W_t^Q H_t^{AC}, K_t = W_t^K H_t^{GA}, V_t = W_t^V H_t^{GA} \tag{6}$$

$$Q_t^{Att} = \mathrm{Relu}(\mathrm{Softmax}(\frac{Q_t(K_t)^T}{\sqrt{d_{att}}}V_t) \tag{7}$$

$$O_t^{Att} = \mathrm{Relu}(\mathrm{Softmax}(\frac{Q_t^{Att}(K_t)^T}{\sqrt{d_{att}}}V_t) \tag{8}$$

where W_t^Q, W_t^K and W_t^V are trainable weight matrix, the division of the similarity matrix and $\sqrt{d_{att}}$ in (7) and (8) contribute to steady gradient descent while training.

2.4. Multi-task Training and Stream/non-stream Decoding

During the multi-task accent ASR model training, the overall loss function is designed by combining three losses: connectionist temporal classification (CTC) loss from ASR task, decoder attention loss and the accent identify cross entropy (CE) loss, which can be formulated as:

$$\mathcal{L}_{all} = \mathcal{L}_{att} + \lambda_{ctc}\mathcal{L}_{ctc} + \lambda_{aid}\mathcal{L}_{aid} \tag{9}$$

$$\mathcal{L}_{ctc} = \mathrm{CTC}(O_t^{Att}, y_f) \tag{10}$$

$$\mathcal{L}_{att} = \mathrm{CE}(\mathrm{Decoder}(O_t^{Att}, y_c), y_c) \tag{11}$$

$$\mathcal{L}_{aid} = \mathrm{CE}(\hat{y}_t^{ac}, y_{ac}) \tag{12}$$

where λ_{ctc} and λ_{aid} are two weights of CTC loss and AID loss. $\hat{y}_t^{ac}$ and y_{ac} are the predicted and true accent label of the input X_t. y_c and y_f are the transcription labels with coarse-grained and fine-grained units. $\mathrm{CE}(\cdot)$ and $\mathrm{Decoder}(\cdot)$ stands for the cross entropy loss and attention decoder function.

In this paper, we use the dynamic chunk masking strategy [6] to ensure compatibility of model inference with both stream and non-stream modes. During the training stage, we initially sample a random chunk size C from a uniform distribution ranging between 1 and the maximum batch length T. Subsequently, the input is divided into multiple chunks based on the selected chunk size. Finally, in training, the current chunk undergoes bidirectional chunk-level attention with itself and previous/following chunks through left-to-right and right-to-left attention decoder respectively.

3. Experiments

3.1. Dataset Description

In this study, we conducted extensive experiments on KeSpeech [8], which involved 1,542 hours of speech signals recorded by 27,237 speakers from 34 cities across China. The dataset encompasses standard Mandarin and its eight subdialects in the regions of Zhongyuan, Southwestern, Ji-Lu, Jiang-Huai, Lan-Yin, Jiao-Liao, Northeastern and Beijing. The MagicData-RAMC, which contains 180 hours and 6 diverse domains (Sichuan, Shanxi, Shandong, Jiangsu, Hunan, Guangdong), is also used as performance validation [26].

3.2. E2E Based Basedline

For [9, 27] acoustic feature extraction, the 80-dimensional log Mel-filter bank (FBANK) is calculated with window size of 25ms and step size of 10ms. The utterance-level cepstral mean and variance normalization (CMVN) calculated using the training set was applied to FBANK for feature normalization. All our experiments are implemented using Wenet end-to-end speech processing toolkit. SpecAugment [28] is used for data augmentation during training, and no extra language models are applied.

3.3. Layer-adapted Fusion Module

3.3.1. layer-adapted Module

As show in Figure 1, The Frame-level general acoustic embedding is obtain from the general acoustic encoder (h_i^{GA}, i stands for the index of layer). The fusion input (H_t^{GA}) is stacked by $\{h_i^{GA}...h_{i+6}^{GA}\}$. In our work, we chose the output of L-6th to L-12th after the progressive downsampling operation as the input for our layer-adapted module. Learnable weights W are introduced into each layer for dot multiplication. The input $W \times H_t^{GA}$ is fused through casual conv2d (kernal size 5x5, stride size 1x1). Accent frame-level predict label is calculated through conv1d of kenerl size 3 and stride size 1.

For AID task, we do some experiments to prove that the different layers of general acoustic encoder contain different-level accent information. We employ a pretrained acoustic encoder, freeze the encoder weights and utilize various layers as inputs to the AID model. Table 1 shows, without using layer-adapted module, L-8th has the highest accuracy for AID on the

Table 2: *Ablation and contrast experiments results of the proposed stream/non-stream Qifusion-Net on the KeSpeech dataset.*

ID	Model	AID	ASR WER(%)									
		ACC(%)	Total	Beijing	Ji-Lu	Jiang Huai	Jiao Liao	Lan Yin	Mandarin	North eastern	South western	Zhong yuan
C1	KeSpeech Baseline	61.13	10.38	11.5	11.5	15.9	11.7	11.7	6.1	10.2	11.9	9.6
C5	DIMNet w/o LM	78.57	9.40	-	-	-	-	-	-	-	-	-
C6	DIMNet w/ LM	78.57	8.87	-	-	-	-	-	-	-	-	-
A1	Qifusion-Net w/o lam	-	15.35	17.13	17.59	23.63	17.93	19.43	8.32	11.69	16.87	14.19
A2	Qifusion-Net fusion-sum	73.41	9.03	10.93	10.4	14.41	10.84	10.12	4.75	8.99	10.68	7.81
A3	Qifusion-Net w/o cif	76.46	9.6	11.25	11.11	15.13	11.24	10.89	5.16	8.99	11.23	8.46
A4	Qifusion-Net self-att	**79.10**	8.25	9.32	9.37	13.22	9.92	9.32	4.57	**7.77**	9.28	7.37
Q1	Qifusion-Net-s	76.54	8.9	10.43	10.18	14.33	10.49	10.08	4.75	8.76	10.24	7.85
Q2	Qifusion-Net-ns	**79.10**	**8.08**	**9.71**	**9.15**	**12.8**	**9.58**	**9.18**	**4.45**	8.28	**9.27**	**7.12**

test set. From L-6th to L-8th, the accuracy increases as the layers deepen, but it declines to a certain extent after L-8th. The accuracy of the AID task with the layer-adapted module reaches 79.1%, which exceeds the absolute improvements of both the KeSpeech baseline [8] (17.91%) and DIMNet [14] (0.28%).

We also investigated the performance differences between concatenate and weight sum connections in layer-adapted architectures. As shown in Table 2, when compared to concatenate (Q2), weight sum (A2) exhibited a 10.5% relative increase in CER performance."

3.3.2. Cross-attention Module

Before the linear classification layer in AID task, we obtain the frame-level accent embedding (H_t^{AC}). The (H_t^{AC}) and the last-layer of general acoustic encoder (h_{12}^{GA}) have the same temporal resolution and feature dimension 256. $Q_t^{Att} = CrossAtt(H_t^{AC}, h_{12}^{GA})$. Q_t^{Att} is used as the attention decoder input with the accent bias eliminated. The attention decoder uses a bi-directional 3-layer transformer structure with a multi-head of 4.

For the AID-ASR MTL system, we do ablation experiments to discuss the impact of different modules on the overall CER. Without AID task, the overall CER reached 15.35 (A1). Added AID task, but without cross-attention module, CER increased to 9.6 (A3). Joint training has a significant positive effect on CER of ASR task. For accent information, cross-attention (Q2) achieves an absolute CER improvement of 0.17 compared to self-attention module (A4). It shows that the addition of cross-attention module can effectively eliminate the degradation of recognition caused by accent in ASR task.

3.3.3. Stream/non-stream Decoding

Based on the dynamic chunk masking strategy, the model supports steaming decoding mode. Compared with baseline and the sota model DIMNet, without LM, the CER of non-stream model (Qifusion-Net-ns Q2) is 10% higher than DIMNet without LM (C5) and even exceeds the DIMNet with LM (C6). The stream decoding mode (Qifusion-Net-s Q1) also outperforms C5 in terms of CER, reaching 8.9. This enables Qifusion-Net-s to match the recognition accuracy of sota model in streaming multi accent system scenarios, which has great potential in practical applications.

Table 3: *Comparison the CER of different models on the MagicData dataset*

Model	AID ACC(%)	RMAC/test CER(%)
LAS-Conformer [26]	-	19.1
Conformer-ASR [29]	-	18.6
CVAE TT [30]	-	17.6
Qifusion-net-ns	**82.8**	**16.3**

Table 3 presents the experimental result on the Magicdata-RMAC. For the ASR task, the first row represents the official baseline [26], while the second row corresponds to the conformer frameworks baseline [29]. The third row displays the latest CER results without a cross-modal extractor [30]. Finally, our proposed Qifusion-Net-ns achieves outstanding performance with a 17.2% lower CER compared to the baseline and attains an accuracy of 82.8% in the AID task.

4. Conclusion

In this study, we explore an end-to-end asr decoding framework in multi-accent systems without prior accent information. Based on the standard conformer ASR architecture, we propose a Qifusion-Net, AID-ASR multi-task learning method based on shared progressive sub-sampling conformer encoder and layer-adapted fusion. We prove that layer-adapted improves AID task, while cross-fusion is beneficial for ASR tasks. The proposed method has lower CER than baseline and sota results in multi accent datasets.

5. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[2] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," *arXiv preprint arXiv:2206.08317*, 2022.

[3] X. Chang, T. Maekaku, P. Guo, J. Shi, Y.-J. Lu, A. S. Subrama-

nian, T. Wang, S.-w. Yang, Y. Tsao, H.-y. Lee *et al.*, "An exploration of self-supervised pretrained representations for end-to-end speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 228–235.

[4] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[5] H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. E. Y. Soplin, T. Hayashi, and S. Watanabe, "Espnet-st: All-in-one speech translation toolkit," *arXiv preprint arXiv:2004.10234*, 2020.

[6] B. Zhang, D. Wu, Z. Peng, X. Song, Z. Yao, H. Lv, L. Xie, C. Yang, F. Pan, and J. Niu, "Wenet 2.0: More productive end-to-end speech recognition toolkit," *arXiv preprint arXiv:2203.15455*, 2022.

[7] Z. Gao, Z. Li, J. Wang, H. Luo, X. Shi, M. Chen, Y. Li, L. Zuo, Z. Du, Z. Xiao *et al.*, "Funasr: A fundamental end-to-end speech recognition toolkit," *arXiv preprint arXiv:2305.11013*, 2023.

[8] Z. Tang, D. Wang, Y. Xu, J. Sun, X. Lei, S. Zhao, C. Wen, X. Tan, C. Xie, S. Zhou *et al.*, "Kespeech: An open source speech dataset of mandarin and its eight subdialects," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[9] X. Wang, Y. Long, Y. Li, and H. Wei, "Multi-pass training and cross-information fusion for low-resource end-to-end accented speech recognition," Jun 2023.

[10] H.-J. Na and J.-S. Park, "Accented speech recognition based on end-to-end domain adversarial training of neural networks," *Applied Sciences*, vol. 11, no. 18, p. 8412, 2021.

[11] H. Hu, X. Yang, Z. Raeesy, J. Guo, G. Keskin, H. Arsikere, A. Rastrow, A. Stolcke, and R. Maas, "Redat: Accent-invariant representation for end-to-end asr by domain adversarial training with relabeling," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6408–6412.

[12] N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. H. Chau, "Best of both worlds: Robust accented speech recognition with adversarial transfer learning," *arXiv preprint arXiv:2103.05834*, 2021.

[13] J. Luo, J. Wang, N. Cheng, E. Xiao, J. Xiao, G. Kucsko, P. O'Neill, J. Balam, S. Deng, A. Flores *et al.*, "Cross-language transfer learning and domain adaptation for end-to-end automatic speech recognition," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[14] Q. Shao, P. Guo, J. Yan, P. Hu, and L. Xie, "Decoupling and interacting multi-task learning network for joint speech and accent recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 459–470, 2023.

[15] Z. Dan, Y. Zhao, X. Bi, L. Wu, and Q. Ji, "Multi-task transformer with adaptive cross-entropy loss for multi-dialect speech recognition," *Entropy*, vol. 24, no. 10, p. 1429, 2022.

[16] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4854–4858.

[17] L. Maison and Y. Esteve, "Improving accented speech recognition with multi-domain training," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[18] L. Ai, S.-Y. Jeng, and H. Beigi, "A new approach to accent recognition and conversion for mandarin chinese," *arXiv preprint arXiv:2008.03359*, 2020.

[19] Y. Qian, X. Gong, and H. Huang, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2842–2853, 2022.

[20] S. Ghorbani and J. H. Hansen, "Advanced accent/dialect identification and accentedness assessment with multi-embedding models and automatic speech recognition," *arXiv preprint arXiv:2310.11004*, 2023.

[21] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning," in *Interspeech 2018*, Aug 2018. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2018-1864

[22] M. T. Turan, E. Vincent, and D. Jouvet, "Achieving multi-accent asr via unsupervised acoustic model adaptation," in *Interspeech 2020*, Oct 2020. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2020-2742

[23] X. Gong, Y. Lu, Z. Zhou, and Y. Qian, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," in *Interspeech 2021*, Aug 2021. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2021-1075

[24] A. Jain, V. P. Singh, and S. P. Rath, "A multi-accent acoustic model using mixture of experts for speech recognition," in *Interspeech 2019*, Sep 2019. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2019-1667

[25] M. Burchi and V. Vielzeuf, "Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2021. [Online]. Available: http://dx.doi.org/10.1109/asru51503.2021.9687874

[26] Z. Yang, Y. Chen, L. Luo, R. Yang, L. Ye, G. Cheng, J. Xu, Y. Jin, Q. Zhang, P. Zhang, L. Xie, and Y. Yan, "Open source magicdata-ramc: A rich annotated mandarin conversational(ramc) speech dataset."

[27] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[28] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, Sep 2019. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2019-2680

[29] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[30] K. Wei, B. Li, H. Lv, Q. Lu, N. Jiang, and L. Xie, "Conversational speech recognition by learning audio-textual cross-modal contextual representation," Oct 2023.